DOCUMENT RESUME

ED 235 234 TM 830 660

AUTHOR Jolly, S. Jean
TITLE Combining Standardized Achievement Test Items with Locally-Developed Supplementary Items to Produce Objective-Referenced Results for Program Evaluation: A Validation Study.
PUB DATE Apr 83
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administrators; *Criterion Referenced Tests; Data Analysis; *Evaluation Methods; Intermediate Grades; *Mathematics Curriculum; Measurement Techniques; Norm Referenced Tests; Program Effectiveness; *Program Evaluation; Teachers; *Test Construction; Test Items; *Test Validity
IDENTIFIERS Stanford Achievement Tests

ABSTRACT

A procedure was developed to combine subsets of curriculum-relevant items from a norm-referenced achievement test (NRT) with items comprising a locally-developed supplementary test in order to produce objective-referenced test (ORT) data for program evaluation. The study was designed to determine whether the ORTs were valid instruments for program evaluation. This was accomplished by analyzing test data gathered from students, classroom performance indices obtained from teachers, and questionnaire responses gathered from school-level administrators. Overall, the results of the study reveal that the ORTs are valid instruments for evaluating the intended programs. Findings indicate that combined NRT and supplementary items are internally consistent, that student ORT scores are related moderately to their classroom performance, and that users perceive the ORTs to be appropriate for evaluating their school-level programs. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED235234

Combining Standardized Achievement Test Items with Locally-Developed Supplementary Items to Produce Objective-Referenced Results for Program Evaluation: A Validation Study

S. Jean Jolly
Palm Beach County Schools

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. J. Jolly

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the
National Council on Measurement in Education

Montreal
April, 1983

TM 830 660

A procedure was developed to combine subsets of curriculum-relevant items from a norm-referenced achievement test (NRT) with items comprising a locally-developed supplementary test in order to produce objective-referenced test (ORT) data for program evaluation. The study was designed to determine whether the ORTs were valid instruments for program evaluation. This was accomplished by analyzing test data gathered from students, classroom performance indices obtained from teachers, and questionnaire responses gathered from school-level administrators.

Overall, the results of the study reveal that the ORTs are valid instruments for evaluating the intended programs. Findings indicate that combined NRT and supplementary items are internally consistent, that student ORT scores are related moderately to their classroom performance, and that users perceive the ORTs to be appropriate for evaluating their school-level programs.

Combining Standardized Achievement Test Items with Locally-Developed Supplementary Items to Produce Objective-Referenced Results for Program Evaluation: A Validation Study

Until the mid-sixties, the primary reason for administering a norm-referenced achievement testing program was to measure individual student achievement. With the advent of mandates for the evaluation of federally-funded educational programs, however, "large scale testing of students became a tool for program evaluation" (Williams and Bank, 1980, p. 3). Norm-referenced tests have apparent drawbacks as instruments for program evaluation, and educators have come to disparage their use for this purpose (Popham, 1975). Critics generally claim that a test designed to have widespread applicability for assessing individual student differences will not, in most cases, be sensitive to specific program objectives (e.g., Madaus et al., 1979; Popham, 1978). For example, some test content may not match program objectives and/or the test may not measure some program objectives.

According to Airasian (1979), many educators viewed the emergence of criterion-referenced tests as a solution to the measurement problems inherent in norm-referenced testing. Indeed, the motive for introducing these tests was to provide the desired kinds of test scores for objective-based instructional programs (Hambleton et al., 1978).

Test publishing companies have attempted to meet educators' needs by providing criterion-referenced scoring and reporting schemes for norm-referenced tests, as well as norm-referenced scoring and reporting schemes for criterion-referenced tests. The first procedure, however, does not address the problem of program objectives which the test does not measure,

and there is a paucity of research on which to base an assessment of the technical adequacy of the second procedure. Regardless, even if the latter procedure proves to be technically adequate, the problem will not have been solved completely. A customized criterion-referenced test with normative scoring and reporting cabability may be two and one-half times as expensive as a traditional norm-referenced test (CTB, 1982), and may not be suitable in all subject areas. The fact remains, however, that a school testing program generally must provide answers to two questions: 1) In relative terms, how does local student achievement in reading, language, mathematics, social science, and science compare to the national sample, and 2) Specifically, are students learning what is being taught; that is, are instructional programs working?

The Problem and Its Significance

Criticism of norm-referenced tests as instruments for program evaluation appears to be a validity issue, the question being whether valid inferences about program effects can be made from a test lacking "curriculum relevance", which may be defined as the correspondence of a test to a particular set of instructional objectives (Cole and Nitko, 1981). Curriculum relevance is a crucial factor for the evaluator, who must be certain that the test items adequately represent the program objectives.

The purpose of this study was to evaluate a method for producing objective-referenced test data from a norm-referenced test. Essentially, mathematics items on two levels and three forms of a norm-referenced test series were matched with district mathematics objectives. Then supplementary mathematics tests were developed and administered just prior to administration of the norm-

referenced tests. After the latter had been scored and traditional normative data generated, those items which matched district objectives were combined with the supplementary items in order to produce objective-referenced test reports. Evaluation of this methodology centered on one seminal question: Are the Objective-Referenced Test (ORT) results valid for assessing the outcomes of the mathematics programs in grades four through six?

Four basic assumptions guided the data collection and analysis activities. First, it was assumed that the selection of norm-referenced items for the ORT item pool had a sound foundation, resulting in a set of curriculum-relevant items; second, that local items comprising the supplementary tests were both technically sound and curriculum relevant; third, that both the norm-referenced and supplementary tests were administered under standard procedures; and fourth, that the ORT reports were used by school-center administrators.

Ebel (1979, p. 7) maintains that school systems desiring to measure instructional objectives not represented on a standardized, norm-referenced achievement test must "prepare, give, validate, and report the results on a local test that does measure what they have been trying to teach." The significance of this study lies in its potential value for those school districts which cannot afford to pay a publishing company to develop a dual-purpose testing program which generates both norm- and objective-referenced test reports. The method under investigation in the Palm Beach County, Florida, school system may offer a possible solution to this problem.

Definitions

In order to develop a common basis for understanding, selective terms used in this study are defined below.

Objective-Referenced Test. An objective-referenced test is one designed to measure the stated outcomes of instruction, the items reflecting the intent of instruction and matched to those objectives. Objective-referenced test scores reflect the kinds of things an examinee can do.

Norm-Referenced Test. The type of achievement test traditionally termed "norm-referenced" refers to a survey-type achievement test in which content coverage is broad and outcomes spanning several years of instruction are measured. Norm-referenced test scores reflect an examinee's location in a defined group of other examinees.

Program/Curriculum Evaluation. "Program" and "curriculum" evaluation are used interchangeably, defined as the measurement of program objectives for the purpose of program improvement.

Program/Curriculum. "Curriculum" is viewed as an instructional program in the form of a written, officially-adopted set of pre-specified objectives. The terms "program" and "curriculum" are used interchangeably.

Program/Instructional Objective. "Program" and "instructional" objective are used interchangeably, defined as statements of what a group of individuals is expected to do as a result of a program (i.e., intended learner outcomes).

Validity. The validity of a test concerns the appropriateness of an inference drawn from a test score. Validation procedures generally are classified by type (i.e., content, criterion, and construct), each procedure being used to collect a different kind of information relevant to a particular purpose. Methods which can be classified as content validation procedures in education frequently are used to assess the appropriateness of an achievement test for the purpose of evaluating an educational program. These procedures traditionally have been limited to assessing the relevance of test items for a specific purpose, the procedures being employed during

test development. In the present study, however, validation refers to empirical procedures employed after test administration, the procedures being intended to assess the adequacy of test scores for making evaluative judgments about an educational program.

Methodology

Following adoption of the Seventh Edition of the Stanford Achievement Test (SAT/7) in March of 1981, a thorough analysis of the match between SAT/7 mathematics items and the local mathematics objectives was performed (see Bower, 1982). Table 1 contains the results of a re-analysis of that study. The percent of SAT/7 items which matched grade-level mathematics objectives ranged from a low of 58 percent at grade six to a high of 66 percent at grade four, while the percent of grade-level mathematics objectives measured by one or more SAT/7 items ranged from a low of 42 percent at grade four to a high of 69 percent at grade six.

Table 1

Results of SAT/7 Curriculum Match Analysis

Percent of:	Int. I(E) Grade 4	Int. II(E) Grade 5	Int. II(F) Grade 6
SAT/7 items matching grade-level mathematics objectives	66%	61%	58%
Grade-level mathematics objectives measured by 1 or more SAT/7 items	42%	54%	69%

Since the SAT/7 contained many items which did not match the curriculum and the curriculum contained numerous objectives not measured by the SAT/7, the test series was deemed inadequate as an evaluation instrument for the mathematics programs. Our initial plan was to use the locally-developed mathematics item bank to generate tests tailored to the curriculum. The results of a district-wide survey, however, indicated that the majority of teachers and administrators also desired the continued administration of a norm-referenced test. Hence, three Supplementary Mathematics Tests (SMTs) were developed using items from the test item bank, the tests being designed to merge with the SAT/7 items which already had been identified as matching the district mathematics objectives.

The SAT/7 and SMTs were administered in April of 1982 to all students in grades four through six, except full-time special education students. There were 4975 fourth grade students, 5412 fifth grade students, and 5324 sixth grade students who completed both the SAT/7 and the SMT.

When the SMTs were administered, teachers were asked to code the third nine weeks student grade equivalent classification (GEC) on the answer document. Palm Beach County requires that elementary school teachers indicate on each nine weeks report card whether a student is working below-, on-, or above-grade-level in mathematics, and guidelines are provided for this purpose.

A computer program designed to merge selected SAT/7 and SMT items was used to create an ORT item file from which diagnostic reports were generated during the fall of 1982. The file contained 76 fourth grade items, 88 fifth grade items, and 132 sixth grade items. Initially, all SMT items and SAT/7 items previously identified as matching district mathematics objectives were included in the file. Several SMT items subsequently were dropped

because of changes in the mathematics curriculum. Additionally, eight SAT/7 items were removed at grade four and seven at grade five because they had an "omit" (i.e., no response) percent greater than seven. Deletion of these items necessitated the removal of SMT items at grades four and five which measured the same objectives as did the SAT/7 items which were removed. Table 2 shows the final composition of the ORTs, with four items measuring each Unified Curriculum mathematics objective.

Table 2

Composition of ORTs

Grade Level	Number of Objectives	Number of Items		
		SAT/7*	SMT*	Total
4	19	28	48	76
5	22	32	56	88
6	33	57	75	132

*Int. I(E)/SMT 4A - 4th grade
Int. II(E)/SMT 5A - 5th grade
Int. II(F)/SMT 6A - 6th grade

Both individual objective scores and total raw score were reported for students, classes, and school grade levels on locally-designed, computer-generated test report forms. Area mathematics consultants distributed the test results to principals during mid-October to early November of 1982.

This study was designed to substantiate the validity of the ORTs as instruments for evaluating an educational program. Although content validity was assured through selection of curriculum-relevant items, post-administration empirical procedures were employed to provide the necessary support

for validating the ORTs as evaluation instruments. Three validation investigations were conducted, each centered around a research question.

Research Question #1: Is each set of ORT items a homogeneous set?

In order to answer this question, ORT item data were processed through the SPSS RELIABILITY computer program (Nie et al., 1975). In addition to providing the user with coefficients of reliability, the program generates basic summary statistics, including item means, standard deviations, inter-item covariance and correlation matrices, scale means, and point biserial correlations. The ALPHA model was specified for computing the K-R 20 internal consistency estimate for dichotomous data.

Research Question #2: Is there a relationship between student ORT scores and their GECs in mathematics?

The first procedure used to ascertain this relationship was the processing of data through the SPSS PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENTS computer program (Nie et al., 1975). Both the ORT total raw scores and the student GECs (below-, on-, and above-grade-level) were submitted as interval-level variables. Although the GECs were ordinal-level data, empirical evidence generally supports the treatment of such variables as though they conform to an interval-level scale.

The second procedure used to determine the relationship between the ORT scores and the GECs was discriminant analysis, a multivariate procedure for improving the discrimination among groups when there are two or more variables which can be combined. The SPSS DISCRIMINANT computer program (Nie et al., 1975) can, however, operate with only one predictor variable.

The ORT total raw scores were submitted as the predictor variable, and the GECs were input as the criterion variable. Since the maximum number of discriminant functions can be no more than the number of discriminating variables, one function picked up all the information in the discriminating (ORT) variable. The resulting canonical correlation was a measure of association between a single discriminating function and the set of variables which defined group membership.

To this point, the DISCRIMINANT program produced redundant information, the canonical correlation being equal to the eta derived from analysis of variance and the Pearson product-moment correlation, both of which were generated by the PEARSON program. In addition to its analytic component, however, the DISCRIMINANT program produces a classification table which contains the percent of correct and incorrect classifications when the predictor variable is used to classify cases into known groups. The procedure enables one to ascertain whether classification errors tend to fall into certain groups (e.g., the ORT/GEC relationship is stronger in the below-grade-level groups, thereby producing few classification errors).

Research Question #3: Do elementary school principals feel that the ORT results are useful to them in improving the intermediate-grade-level mathematics programs?

After the reports had been disseminated, an ORT School Survey was distributed to the 66 principals of schools containing grades four, five, and/or six. It consisted of four open-ended questions, two relating to perceived problems and two relating to use of the reports. The problem-related questions were included in order to determine whether there might be a negative bias in the "use" responses due to problems encountered in using the reports. The

use-related questions referred to both current and intended use. Additionally, no reference was made to the type of ORT report (i.e., student, class, or school grade-level), although the school grade-level report was the primary focus of interest for this study. Principals were asked to respond to the Survey by December 17, 1982, and a content analysis of their responses was performed during January of 1983.

Results and Discussion

This study was designed to obtain evidence regarding the validity of using ORT data in judging the effectiveness of mathematics programs in grades four through six. The results of the study, which was based on three research questions, are presented below.

Research Question #1: Is each set of ORT items a homogeneous set?

Table 3 contains the K-R 20 coefficients for the ORTs, all of which exceeded 0.90. Given the magnitude of these coefficients, it may be concluded that the ORTs reflect homogeneity of content. Coefficients of internal consistency for achievement tests should exceed 0.90, although estimates of 0.75 for a 50-item test and 0.86 for a 100-item test are sufficient (Ebel, 1980, p. 228).

Table 3

K-R 20 Coefficients for ORTs

Grade Level	Total Number Of Cases	Items	Mean	SD	K-R 20
4	4975	76	56.42	11.68	0.92**
5	5412	88	58.82	14.73	0.93**
6	5324*	132	82.00	24.94	0.96**

*Includes 6th grade middle school students
**p .0001

Research Question #2: Is there a relationship between student ORT scores and their GECs in mathematics?

Tables 4-6 contain the ORT means and standard deviations for the three student GEC groups at each grade level. Sixth grade middle school students were excluded from the analysis because the middle school report card does not contain classifications of below-, on-, and above-grade-level. Additionally, elementary school students for whom no GEC was provided were excluded from the analysis. The differences in the GEC means for each grade level were as expected, with those students whom teachers classified as performing above-grade-level attaining the highest group means and the below-grade-level students evidencing the lowest group means.

Table 4

Means and Standard Deviations for Fourth Grade Student GECs

GEC	N	Mean*	SD
Below-Grade-Level	694	43.45	11.53
On-Grade-Level	3782	58.34	10.13
Above-Grade-Level	167	63.69	8.77

*A test of significance among the means yielded an overall F-ratio significant beyond the .0001 level.

Table 5

Means and Standard Deviations for Fifth Grade Student GECs

GEC	N	Mean*	SD
Below-Grade-Level	1049	44.12	11.76
On-Grade-Level	3723	62.59	12.64
Above-Grade-Level	245	70.54	13.02

*A test of significance among the means yielded an overall F-ratio significant beyond the .0001 level.

Table 6

Means and Standard Deviations for Sixth Grade Student GECs

GEC	N	Mean*	SD
Below-Grade-Level	654	61.01	18.75
On-Grade-Level	2086	91.57	21.17
Above-Grade-Level	306	106.08	22.02

*A test of significance among the means yielded an overall F-ratio significant beyond the .0001 level.

Table 7 contains the Pearson product-moment correlation coefficients for the ORT scores and GEC classifications in grades four through six. The correlations between the ORT scores and the GECs suggest a moderate relationship between the two variables. There are three possible reasons, aside from the degree of relationship between the variables, for the moderate correlations (0.47 to 0.54) which were obtained. 1) These values do not incorporate a correction for attenuation to compensate for the presence of random errors in both variables. 2) Weak to moderately strong correlations in the range of 0.2 to 0.7 are expected when comparing test scores with performance measures (Green, 1981, p. 1002). 3) Since the student GEC variable is trichotomous, a perfect correlation could be obtained only if all test scores at these three points fell exactly on three ORT scale points. This condition is impossible because the three GEC points must correspond to a range of ORT scale points, and the latter variable is continuous and scores are distributed widely.

Table 7

Pearson Product-Moment Correlation Coefficients
Between ORT Scores and GECs

Grade Level	N	r	GEC SD	ORT SD
4	4643	0.47**	0.42	11.68
5	5017	0.52**	0.48	14.73
6	3046*	0.54**	0.55	24.94

*Sixth grade elementary school students only
**p .0001

Tables 8-10 contain the discriminant analysis classification results. There was a relatively high percent of correct classifications for the "below" and "above" groups, ranging from a low of 70 percent of "above" cases at grade five to a high of 80 percent of "below" cases at grade six. The percent of correct classifications for the on-grade-level groups, however, ranged from 33 percent to 39 percent. Nunnally (1978) states that members of different groups should cluster and not be scattered randomly (e.g., if groups are well-discriminated, means are far apart and members hover near the means). Histograms of predicted classifications showed the "on" groups to be scattered across the continuum, from below the "below" group mean to above the "above" group mean. This resulted in an overall correct prediction rate of approximately 45 percent for grades four and five, and 47 percent for grade six. Although these figures are better than chance predictions, the error in the "on" group classification is apparent.

There are two possible reasons, aside from the degree of relationship between the ORTs and GECs, for this low percent of correct classifications.

1) The analyses were conducted as though the GECs contained equal proportions of students, and the Bayesian adjustment, which would have compensated for the disproprtionate groupings, was not employed. 2) It was verified that some teachers had miscoded student GECs.

Table 8

Comparison of Actual and Predicted GECs for Grade Four

GEC	N	Predicted GECs*		
		Below	On	Above
Below-Grade-Level	694	494 (71.2%)	171 (24.6%)	29 (4.2%)
On-Grade-Level	3782	775 (20.5%)	1458 (38.6%)	1549 (41.0%)
Above-Grade-Level	167	15 (9.0%)	28 (16.8%)	124 (74.3%)

*Percent of "grouped" cases correctly classified: 44.71.

Table 9

Comparison of Actual and Predicted GECs for Grade Five

GEC	N	Predicted GECs*		
		Below	On	Above
Below-Grade-Level	1049	831 (79.2%)	178 (17.0%)	40 (3.8%)
On-Grade-Level	3723	887 (23.8%)	1266 (34.0%)	1570 (42.2%)
Above-Grade-Level	245	30 (12.2%)	43 (17.6%)	172 (70.2%)

*Percent of "grouped" cases correctly classified: 45.23.

Table 10

Comparison of Actual and Predicted GECs for Grade Six

GEC	N	Predicted GECs*		
		Below	On	Above
Below-Grade-Level	654	523 (80.0%)	114 (17.4%)	17 (2.6%)
On-Grade-Level	2086	531 (25.5%)	678 (32.5%)	877 (42.0%)
Above-Grade-Level	306	27 (8.8%)	49 (16.8%)	230 (75.2%)

*Percent of "grouped" cases correctly classified: 46.98.

Research Question #3: Do elementary school principals feel that the ORT results are useful to them in improving the intermediate-grade-level mathematics programs?

All ORT School Surveys distributed to elementary school principals were returned, and 49 of the 66 respondents indicated that they already had or intended to use the ORT results for making decisions about the mathematics programs in their respective schools. The comment which seemed most illustrative of the overall response of principals was that the ORT functioned "as a diagnostic tool for program improvement--the only real reason for testing." Table 11 contains the results of a content analysis of the responses which pertained to programmatic uses of the ORT results exclusive of individual student and teacher use.

Table 11

Current and Intended Programmatic Uses of ORT Results

Programmatic use of ORT results:	Frequency
Program review	29
Program planning	21
Articulation among grade levels	14
Selection of program materials	12
Changes in planned activities based on assessment of program strengths and weaknesses	10
Other	24

In addition to using the ORTs for making programmatic decisions, principals indicated that student and class reports would be used for working with students and teachers, respectively. The two Survey questions related to problem areas and suggestions for improvement yielded only one finding: principals asked that ORT results be returned at the end of the school year, rather than six months late. This was not an unexpected finding.

Summary

"Measurement validity concerns the soundness of the inferences that are made from the results of the data-gathering process" (Joint Committee, 1981, p. 116). In order to determine whether the ORTs were valid instruments for program evaluation, three empirical investigations were conducted. The results of these investigations confirmed that the ORTs were homogeneous tests which evidenced moderate correlations with student classroom performance,

and that elementary school principals viewed them as useful in making programmatic decisions.

The Palm Beach County school system will replicate this study during April of 1983, using revised SMTs for grades four through six and newly-developed SMTs for grades two and three. Additionally, plans are underway to implement an ORT program in the area of communication skills.

References

Airasian, P. W. Educational measurement and technological bases underlying minimal competency testing. In P. W. Airasian, G. F. Madaus, & J. J. Pedulla (Eds.), Minimal competency testing. Englewood Cliffs, N. J.: Educational Technology Publications, 1979.

Bower, R. Matching standardized achievement test items to local curriculum objectives. Paper presented at the meeting of the National Council on Measurement in Education, New York, March 1982.

Cole, N. S. & Nitko, A. J. Measuring program effects. In R. A. Berk (Ed.), Educational evaluation methodology: The state of the art. Baltimore, MD: The John Hopkins University Press, 1981.

CTB/McGraw-Hill. Criteria. Monterey, CA, 1982.

Ebel, R. L. Essentials of educational measurement (3rd ed.). Englewood Cliffs, N. J.: Prentice Hall, 1979.

Ebel, R. L. Practical problems in educational measurement. Lexington, MA: D. C. Heath, 1980.

Green, B. F. A primer on testing. American Psychologist, 1981. 36(10), 1001-1011.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48(1), 1-47.

Joint Committee on Standards for Educational Excellence. Standards for evaluation of programs, projects, and materials. New York: McGraw-Hill, 1981.

Madaus, G. F., Kellaghan, T., Rakow, E. A., & King, D. J. The sensitivity of measures of school effectiveness. Harvard Educational Review, 1979, 49(2), 207-230.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical package for the social sciences. New York: McGraw-Hill, 1975.

Nunnally, J. C. Psychometric theory (2nd ed.). New York: McGraw-Hill, 1978.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1978.

Popham, W. J. But . . . The measurement cupboard was bare. CEDR Quarterly, 1975, 8(1), 12-16.

Williams, R. C. & Bank, A. A school district linking of evaluation and testing with instruction: A problem of attention. Paper presented at the meeting of the Evaluation Research Society, Arlington, VA, November, 1980.

209/G
3/15/83
6:00 P.M.
(Pages 1-21)